

Amendments to the Claims

Please amend the claims as follows. The Claim Listing below will replace all prior versions of the claims in the application:

Claim Listing

1. (Currently amended) A method for determining if a first and second document stored in a digital format in a data processing system are similar by comparing sparse representations of the two documents, the method comprising the steps of:

breaking the first and second documents into chunks of data of predefined sizes;
selecting a subset of all chunks as representative of the data in ~~the~~ each document;
determining a set of coefficients that represent the selected chunks in each document;
combining sets of coefficients for each document into coefficient clusters for each document, a coefficient cluster containing coefficients which are similar according to a predetermined similarity metric; and
evaluating a degree of similarity between the two documents by counting coefficient clusters into which chunks from both documents fall.

2. (Original) A method as in claim 1 wherein the coefficients that represent a particular chunk are selected as Fourier transform coefficients for data values that make up the chunk.

3. (Original) A method as in claim 2 wherein the selected coefficients are the absolute values of the Fourier transform coefficients.

4. (Original) A method as in claim 2 in which the data in a chunk is mapped onto a unitary circle in a plane of complex variables before Fourier coefficients are calculated.

5. (Original) A method as in claim 1 wherein a degree of similarity is determined by calculating a correlation of coefficients of the data stored in the chunks.

6. (Currently amended) A method is as in claim 5, in which the correlation is linear, after outliers are removed from the ~~vectors~~ sets of coefficients.
7. (Original) A method as in claim 1 wherein the step of evaluating a degree of similarity is carried out in a manner to account for possible shifts in the position of similar data in the two documents.
8. (Currently amended) A method as in claim 1 wherein the ~~cluster representation comprises~~ coefficient clusters are represented as a hierarchy having at least two levels, where successively lower levels of the hierarchy represent only portions of the chunks at higher levels of the hierarchy.
9. (Currently amended) A method as in claim ~~[[1]]~~ 8 wherein the step of ~~comparing~~ evaluating proceeds first at a higher level in the hierarchy, and if a ~~sufficient~~ predetermined degree of similarity between coefficients of a ~~queried~~ chunk and centers of the clusters is found at the higher level, only then proceeding to compare coefficients at a lower level in the hierarchy.
10. (Currently amended) A method as in claim 9 wherein the ~~comparison~~ evaluation of coefficients of chunks to clusters at a given lower level in the hierarchy is limited to consideration of only the clusters belonging to those branches of the hierarchy which run through related higher-level clusters already determined to be similar to the ~~coefficients of the queried~~ document.
11. (Currently amended) A method as in claim 9 further comprising:
 - a. selecting a cluster exploration set derived from a set of coefficients located at a predetermined level of the hierarchy for the first document;
 - b. computing a similarity for clusters in the cluster exploration set, by comparing the clusters in the cluster exploration set against at least one chunk of the second document selected as a base element;
 - c. sorting the clusters so compared according to their degree of similarity to the

- chunk from the second element;
- d. calculating a penetration similarity threshold;
 - e. selecting a subset of the cluster exploration set as those clusters that are ~~most~~ similar to the base element;
 - f. treating ~~this~~ the subset further as a next cluster exploration set; and
 - g. repeating steps b to f until ~~[[a]]~~ the bottom of the hierarchy is reached; and
 - h. returning the subset generated at step f as the solution otherwise.
12. (Original) A method as in claim 1 wherein the step of comparing further comprises: a query interpretation process for merging results of queries for multiple chunks in the hierarchy, to determine an overall degree of similarity for the two documents.
13. (Original) A method in claim 12 additionally wherein the first document is determined to be similar to a group of documents within a larger set of pre-processed documents by the further step of:
- determining a number of similar chunks within the first document and all documents in the set of pre-processed documents that have been pre-processed by the method.
14. (Original) A method in claim 13 wherein documents in the set of pre-processed documents which have fewer than a predetermined number of chunks similar to the first document, are not considered to be similar.
15. (Original) A method as in claim 11 wherein out of a subset of clusters generated at step f, a cluster which, together with its parent upper-level clusters of the hierarchy is most similar on average to a given set of coefficients is selected as a host for storing the corresponding set of coefficients.
16. (Original) A method as in claim 15 in which an average similarity of clusters at different levels of the hierarchy to the corresponding set of coefficients is given by an arithmetic average of squares of similarities of clusters at different levels with the said set of coefficients, weighted

by a dimension of clusters at those levels.